May 21, 2013

The Statistical Properties of the Intrinsic Estimator for Age-Period-Cohort Analysis

Yang Claire Yang, University of North Carolina at Chapel Hill

Kenneth C. Land, Duke University

Introduction

Luo (2013) engages in a discussion and critique of the Intrinsic Estimator (IE) of the age, period, and cohort (APC) components of temporally ordered arrays of age-graded population rates or proportions in the context of the classical APC multiple classification/accounting model. The specific subject matter of Luo (2013) is an assessment of "… the validity and application scope of…" of the IE. The assessment is based on theoretical and simulation analyses. Unfortunately, Luo (2013) reiterates many statements about the IE that have been acknowledged for years, makes erroneous interpretations, claims, and assertions about the IE, and identifies and analyses a specific instance in which the IE should never be applied. This article identifies and clarifies these statements and errors and the fundamental limitations of Luo's (2013) analyses.

The Underidentification Problem of the Classical APC Accounting/Multiple Classification Model

The classical Age-Period-Cohort Accounting/Multiple Classification Model for an age-by-time period table of population rates or proportions (with cohorts defined by birth or another shared event arrayed in the diagonals of the tables) was stated algebraically over 40 years ago by Mason, Mason, Winsborough, and Poole (1973) for a classical normal errors regression model and for which there are analogous Generalized Linear Model specifications for non-normally distributed outcomes.[1] With a set of identical fixed widths of the age and time period intervals, this model is underidentified – the $(X^TX)^{-1}$ inverse matrix (a function of the design matrix X) used in estimation of the coefficient vector b of the model is deficient by the

[1] The standard notation for this model is specified in Equations (1) and (2) of Luo (2013) and will not be repeated here.

rank of one. Accordingly, one constraint, such as an equality constraint on two of the coefficients of *b* is sufficient to identify the model and estimate the vector of age, period, and cohort coefficients – for example, setting two age group, time period, or cohort coefficients equal to each other. This then produces a Constrained Generalized Linear Model (CGLIM)[2] for the APC analysis. As Luo (2013) notes, various students of this model (e.g., Rodgers 1982; Yang et al. 2004; Glenn 2005) have shown that CGLIM estimates of the coefficients of a constrained coefficient vector are sensitive to the choice of equality constraints.

Suffice it to say that the underidentification problem of the classical APC accounting model has been extensively studied over the past four decades. We believe that most students of this model and the identification problem agree that the coefficient vector *b* is not identified and cannot be estimated from the data (see, e.g., the discussions by O'Brien 2011a; 2011b and Fu, Land, and Yang 2011). That is, only constrained coefficient vectors can be estimated.[3] Both theoretical analyses and simulation studies have shown that, if the equality constraint imposed on the coefficient vector *b* of the APC accounting model is the "true" constraint in the sense that it is satisfied by the underlying parametric model, then statistical estimates of the coefficients of the constrained vector will accurately estimate the underlying coefficient vector (within sampling error) that generated the data. Given the sensitivity of estimates of constrained coefficient vectors to the equality constraint imposed, however, the question always emerges: How can we be sure that the equality constraint imposed is a "true" constraint in the sense that it is satisfied by the "generating mechanism" that produced an observed matrix of rates? This question has motivated numerous articles on the APC

[2] We follow Yang et al. (2004; 2008), using CGLIM for this class of equality-constrained models, as the GLIM abbreviation for Generalized Linear Models dates back to McCullough and Nelder (1983); the abbreviation GLM also is used for Generalized Linear Models, a convention which Luo (2013) follows.
[3] Fu et al. (2011) compared the identification problem in the APC accounting model to the problem of defining contrasts that can be estimated in a classical analysis of variance model.

accounting model, it is the question that motivates Luo's (2013) attempt to assess the validity of the Intrinsic Estimator and the constraint on the unidentified coefficient vector of the APC accounting model in this context, and it is the question that is the source of much of the confusion and erroneous statements therein.

The Coefficient Vector Estimated by the Intrinsic Estimator

Luo (2013) makes erroneous statements about our claims regarding the coefficient vector estimated by the IE and for which the IE is unbiased. This is amazing as we have stated this very clearly in our various articles on the IE, for example, in Yang, Fu, and Land (2004:101): "Theorem 1: The Intrinsic Estimator is an unbiased estimator of $[b_1]$ (using the notation adopted here rather than that of Yang et al. 2004) in finite-time-period APC analyses of any fixed number p of time periods."[4] Let this be reiterated and clearly stated yet again: *The Intrinsic Estimator does not estimate the unconstrained coefficient vector of the APC accounting model. It estimates the projection of the unconstrained vector onto the non-null subspace of the vector space defined by the columns of the design matrix of the accounting model.* While the use of language in various descriptions of the IE and its properties may

[4] In the section on Biasedness, Luo (2013) states the definition: "By definition, an estimator δ is an unbiased estimator of a parameter θ if the expectation of δ over the distribution that depends on θ is equal to θ, or $E_\theta(\delta) = \theta$." This is a correct, conventional definition of unbiasedness as a property of a statistical estimator. In footnote 6, Luo (2013) then asserts: "Yang and colleagues have used "unbiasedness" in a different sense: they used this term to mean that the expectation of IE is equal to b_1, the projection of parameter vector b onto the null space of design matrix X (e.g., see *ibid*. p 1709). This is an important distinction because the true parameter vector b can be very different from its projection onto the non-null space b_1, the vector that IE actually estimates. Because APC analysts are usually interested in estimating the true age, period, and cohort effects, the classic concept of unbiasedness is more relevant to APC research than that used by the IE proponents. Thus I use "unbiasedness" in its classic sense for the following discussions." This is a distortion of the analyses of Yang et al. (2008) referenced in Luo's (2013) footnote 6. It is correct that Yang et al. (2008) used the property of unbiasedness to "mean that the expectation of IE is equal to b_1, the projection of parameter vector b onto the null space of design matrix X" precisely because that is the constrained parameter vector estimated by the IE. And this usage is entirely consistent with Luo's (2013) quoted definition of the unbiased property. The question of whether an estimate of the projected coefficient vector b_1 is a "good" or "unbiased" estimate of "the true age, period, and cohort effects" is a separable question and has nothing to do with the unbiasedness property of the IE as an estimator of the coefficient vector it estimates.

have been sufficiently informal that this has not always been clearly conveyed, and while some researchers may not have completely understood this property, we have been clear from the outset of our works about this matter.

What is the Intrinsic Estimator?

In the Conclusion and Discussion section, Luo (2013) states: "… IE is nothing new in APC analysis. Kupper and his colleagues were the first to introduce the IE solution to APC analysts; they referred to this solution as the Principal Component Estimator (PCE) (Kupper et al. 1983, p2795-97)." This is incorrect. First, the Principal Component (PC) Estimator is not the same thing as the IE. The IE transforms the PC regression coefficient estimates back to the original coordinates of the APC accounting model so that they can be interpreted in terms of those coordinates; geometrically, this can be interpreted as projecting the PC regression coefficients estimates to the hyperplane in the coefficient space defined by the APC accounting model vector *b*. This hyperplane is defined by the IE constraint $Xb_0 = 0$. This transformation of the PC regression coefficients (those estimated by Kupper et al. 1985) back to the original coordinates of the APC accounting model.

This is a unique contribution of the IE that was recognized by Fu and described in the various publications on the IE with his co-authors. This transformation makes the IE estimates more useful substantively than the PC coefficients. Perhaps the description of the IE estimated A, P, and C coefficients geometrically as a projection to the hyperplane in the coefficient space defined by the IE constraint is difficult to grasp, but other ways to describe the same thing should be used. While it is true, as O'Brien (2011a) states and as cited by Luo (2013), that the PC and IE coefficients are algebraically equivalent, the rescaling facilitates

substantive interpretation, and the choice of scale is often of key importance for substantive interpretation in many areas of scientific studies and statistical analysis.

Second, the key article that led to the conceptualization and subsequent development of the IE is Fu (2000). This article showed that *the IE is a special case of the classical ridge estimator for the conventional linear regression model that is used when the regressors are highly collinear*. Fu (2000) studied the ridge estimator in the singular design case, where the design matrix X has one-less than full rank, of which the design matrix for the APC accounting model is an example, and showed: 1) that the ridge estimator lies in a sub-parameter space orthogonal to the null space of the design matrix generated by the eigenvector of the zero eigenvalue, and 2) that the ridge estimator converges to the IE as the shrinkage parameter λ tends to 0. In other words, the Intrinsic Estimator can be interpreted as the limit of the ridge estimator as its shrinkage penalty goes to zero.

In brief, Kupper et al. (1985) laid a lot of methodological groundwork for the IE, but they did not define it. Fu's initial contributions were (1) to introduce the rescaling of the PC regression coefficients back to the original coordinates of the APC model, and (2) to demonstrate the relationship of the IE to the ridge estimator. These contributions were elaborated, studied, and empirically applied by Fu and his coauthors (Yang, et al. 2004; 2008) in subsequent publications.

More recently, Tu, Kramer, and Lee (2012) studied the application of partial least squares (PLS) to the APC accounting model. Whereas principal components regression extracts the components independently of the outcome variable, PLS maximizes the covariance of the components with the outcome variable Y, extracting the components by order of this covariance from the highest to the lowest. Tu et al. (2012) showed that, as the

number of components extracted by PLS approaches the maximum number possible for a design matrix, the numerical values of the PLS estimates of the age, period, and cohort effect coefficients approach, and are within sampling error of, the corresponding coefficients estimated by the IE (which uses the maximum possible number of components). They also showed that an estimator based on the first three PLS components is a numerically reasonable approximation to the PLS effect coefficients estimated by using the maximum possible number of components. This lends yet additional insight into the nature of the Intrinsic Estimator and its robustness.

The Linear Constraint Applied to Obtain the Intrinsic Estimator

The section on "The Linear Constraint Implied by IE" of Luo (2013) includes an algebraic derivation of "the specific form of this constraint for data sets with varying numbers of age, period, and cohort groups." The derivation appears to be technically correct; however, it is based on the assumption stated in the sentence "To illustrate, suppose that age, period, and cohort each have effects on the outcome variable that show a linear trend." Very simply and straightforwardly, *this is a case in which the IE should never be applied to estimate age, period, and cohort effect coefficients – a case in which all three of the age, period, and cohort groups have exact algebraic linear trends*. The reason is very simple: In this case, any one of the three can be written as an exact linear function of the other two – which is what is demonstrated in this section of Luo (2013).

In empirical applications of the IE, for example, those in Yang et al. (2004) and Yang (2008), it always has been emphasized that a researcher should conduct preliminary model specification tests, using, for example, the AIC or BIC models selection statistics, prior to

specifying the form of the APC model to be estimated. If these model selection tests indicate that one or two of the three temporal dimensions of the APC model are sufficiently collinear with the other dimensions that they do not contribute significantly to the outcome variable, then the analyst should not specify the full three-way APC model but rather a reduced model with one or two of the temporal variables.[5] The reason is that the IE is an estimator for an APC accounting model in which all three of the temporal dimensions contribute independently to the outcome variable, and, if this is not the case, then the model should not even be specified as the full APC model.

Accordingly, again, *the IE should not be applied when there is an exact linear relationship among the three dimensions.* And adding additional time periods to the tabulated data, as in the Appendix Figure 1 of Luo (2013), will not resolve this problem. In sum, what Luo (2013) has done in this section of the paper is demonstrate the implications of a situation in which the IE should not be applied. Works as early as Yang (2008) and as recent as Yang and Land (2013: Chapter 5) have laid out a three-step procedure that should be thoroughly applied to APC analysis using the accounting model. It is so important that we believe it is worth repeating here. Step 1 is to conduct descriptive data analyses using graphics, with the objective being to provide qualitative understanding of patterns of temporal variations. Step 2 is model fitting and calculation of model fit statistics such as the Bayesian Information Criterion (BIC). The objective is to ascertain whether the data are sufficiently well described by any single factor or two-factor model of age (A), time period (P), and cohort (C) effects for which there is no identification problem. Only when these analyses suggest that all three

[5] Similar remarks apply to the second model specification studied in this section of the paper, that in which all three of the temporal dimensions have both linear and quadratic trends across the effect coefficients. This again is a situation in which all three of the dimensions are not contributing independently to variation in the outcome variables and for which are reduced model should be specified and estimated.

dimensions are operative should one proceed with Step 3: a three-factor APC model to which a constrained estimator can be applied to identify the A, P, and C effects. By revisiting Glenn's (2005) numerical example, Yang and Land (2013: 109) emphasized that "imposition of a full APC model on data when a reduced model fits the data equally well or better constitutes a model misspecification and should be avoided." Empirical examples of chronic disease mortality in Yang (2008) and cancer mortality in Yang and Land (2013) show the necessity of all three steps, whereas those of cancer incidence for certain sites in the latter show the first two steps suffice. A blind application of the IE, or any other constrained estimator, of the full three-factor APC model was never recommended.

The foregoing statements are based on mathematical analysis. In addition, we conducted a statistical analysis of the three models used in the simulations reported in Luo's (2013) Table 3. For this, we first calculated the numerical expected values of the simulations as described by Luo (2013) for the three models. We then applied steps 1 and 2 of the aforementioned three-step procedure. The results of the calculation of the BIC model selection statistics in step 2 are given in Table 1. As the BIC model selection guidelines are that the model with the smallest BIC should be chosen (Raftery 1995), the values of these statistics in Table 1 clearly show that the data are well described by two-factor models with period and cohort effects (Datasets 1 and 2) and age and cohort effects (Dataset 3), respectively. In other words, as expected on the basis of mathematical analysis, the two-factor models show superior model fits to the data generated by the simulations to full, three-factor APC models, and therefore estimation of a three-factor model is completely erroneous and specious.

Table 1. Numerical Values of Bayesian Information Criterion (BIC) Model Selection Statistics for Models Generating the Three Datasets of Luo's (2013) Table 3.

Model	Dataset 1	Dataset 2	Dataset 3
Age	1400.12	10310.12	1310.12
Period	-39.8798	8870.12	1550.12
Cohort	1347.734	1347.734	147.7335
Age and Period	-62.2665	8847.734	1287.734
Age and Cohort	1325.347	1325.347	-114.653
Period and Cohort	-114.653	-114.653	125.3468
Age, Period, and Cohort	-50.8465	8859.153	1299.153

Next, let's consider the question raised in Luo (2013): "What does equation (8) mean?" where equation (8) is:[6]

$$b \cdot b_0 = 0$$

As noted in Luo (2013), this matrix algebraic equation indicates that the projection of the unidentified APC accounting model coefficient vector *b* on the vector corresponding to the projection of *b* onto the non-null subspace of the column vector space of the design matrix *X* is zero. The text of the article also correctly notes that equation (8) is another way of stating the constraint on the unidentified APC coefficient vector that is applied by the IE, namely that the coefficient of the null vector, *s*, equals zero. And, thirdly, equation (8) is another

[6] In fact, equation (8) in Luo is erroneous, as the b_0 vector must be transposed in order to be conformable to the *b* vector so that the product exists and equals zero.

way of stating that the coefficient vector estimated by the IE is $b_0 = P_{proj} b$, that is, the projection of *b* onto the non-null subspace. These three are algebraically equivalent.

Here is the problem: In this section of the paper and elsewhere, equation (8) is termed the "implicit constraint" or "implicit LC [linear constraint]" imposed by the IE. What is "implicit" about it? The constraint has been stated and articulated in expositions of the IE by Fu and coauthors in all expositions of the IE. The use of the term "implicit" by Luo (2013) gives the impression that Fu and coauthors sought to ignore or not explicate this constraint. Nothing could be further from the truth.

Estimating Inestimable Effect Coefficient Vectors

The section on Application Scope in Luo (2013) reports a simulation study for models that have quadratic trends in their age, period, and cohort effect coefficients as specified in four equations. The second, third, and fourth of these models contain only pairs of the age, period, and cohort temporal dimensions rather than all three, and the IE is applied to estimate the coefficient vectors for these reduced models. As pointed out above, this is an inappropriate application of the IE – this estimator should never be applied to data that have been generated by a reduced model or for which model specification tests indicate that this is the case. Accordingly, the results of the IE applications reported in Scenarios 2, 3, and 4 of Figure 1should not be interpreted – they are meaningless.

The first of the models studied in this group contains quadratic trends in all three temporal dimensions.[7] The results of the application of the IE to this model for 9 age groups,

[7] As noted in the preceding section, the IE is an estimator for an APC accounting model in which all three of the temporal dimensions contribute independently to the outcome variable. When all three temporal dimensions have identical algebraic trends across the effect coefficients, as in this quadratic trends specification, model identification tests likely would imply that a model to be estimated should include only two of the three

50 time periods, and 57 cohorts are shown graphically in Figure 1 of Luo (2013). These results are consistent with what we know from both mathematical analysis and other simulation studies about the consistency properties of the IE. First, as Fu has shown in mathematical studies of the asymptotic properties of the IE (Fu and Hall 2006), it is the age dimension for which the effects of the design matrix, which are very large in this simulation study, on the coefficient estimates wear off the fastest. That is, the Age graph in Figure 1 show that the plots of the curves for the IE estimates of the age coefficients are quite close to those of the "true model." Second, because this particular simulation study specified a generating model that has very large effects of the design matrix, the plots of the curves for the IE estimates of the period and cohort coefficients show quadratic patterns that are not as close, with the curve for the cohort effects being closer to that of the generating model than that of the period effects. Again, this is consistent with Fu's asymptotic studies, as the number of cohorts is larger than the number of time periods.

Model Cross-Validation of IE Estimates

Footnote 7 of Luo (2013) makes the following comment about model cross-validation:

> Yang and colleagues have used empirical data, where the true effects are unknown, to assess these properties and performance of IE (see *ibid.*, p1712-16). However, it is logically impossible to assess the performance of an estimator when the true underlying effects are unknown. If such a cross-model validation of the IE for a specific empirical data set were to show that IE yields reasonable estimates, this can only depend on having selected examples that are consistent with the IE's constraint. Therefore, cross-model comparisons using empirical data are not an appropriate method to validate IE.

---

dimensions. Thus, even though the IE likely is not an appropriate estimator for this particular set of simulated data, with a large number of time periods and cohorts (50 and 57, respectively, in this simulation), the asymptotic properties of the IE studied by Fu are evident.

Cross validation of empirical findings by application of alternative statistical models is a fundamental principle of analysis of robustness in applied statistics. And the "underlying true effects" are never known, because one can never know with certainty that the statistical model specified is the "true model." The celebrated statistician George E. P. Box (1979:202), Emeritus Professor of Statistics at the University of Wisconsin at Madison is famous for his statement that: "All models are wrong, but some are useful." In the context of APC analysis, the implication of this statement is that all statistical models are wrong, including both the CGLM and IE models as well as others such as the APCC, HAPC-GLMM, and causal mechanism-based models. However, when we find estimates of age, period, and cohort trends that show some consistency across models, this is indicative that the estimates produced by a specific model are not spurious and increases both our confidence in the empirical findings and our assessment of the usefulness of the model.

Conclusion

Luo (2013) claims: 1) that there is nothing new about the Intrinsic Estimator for the age, period, and cohort effect coefficients of the classical APC accounting model; 2) the IE is not an unbiased estimator of the unidentified coefficient vector of this model; 3) the constraint imposed by the IE on the unidentified coefficient is "implicit"; and 4) the IE performs poorly as a statistical estimator of the unidentified coefficient vector when that vector has very large effects of the design matrix. In the foregoing sections, we have responded that point 1) represents a misunderstanding of the IE, point 2 is a claim that we never made, point 3) is a mischaracterization of our work, and point 4) disregards the asymptotic properties of the IE. In short, there is little of merit in the Luo (2013) paper other than an algebraic demonstration

of a situation – identical linear or nonlinear algebraic trends in the effect coefficients for all three temporal dimensions – in which the Intrinsic Estimator should never be applied.

References

Box, G. E. P. 1979. Robustness in the strategy of scientific model building. In *Robustness in statistics: Proceedings of a workshop*, ed. R. L. Launer, and G. N. Wilkinson, 201-236. New York: Academic Press.

Fu, W. J. 2000. Ridge estimator in singular design with application to age-period-cohort analysis of disease rates. *Communications in Statistics - Theory and Methods* 29:263-278.

Fu, Wenjiang J. and Peter Hall. 2006. "Asymptotic Properties of Estimators in Age-Period-Cohort Analysis." *Statistics & Probability Letters* 76 (17):1925-1929.

Girosi, F. and G. King. 2008. *Demographic Forecasting*. Princeton: Princeton University Press.

Glenn, Norval D. 2005. *Cohort Analysis*. Second Edition. Thousand Oaks, Calif.: Sage Publications.

Kupper, Lawrence L., Joseph Janis, Ibrahim A. Salama, Carl N. Yoshizawa, Bernard G. Greenberg, and H. H. Winsborough. 1983. "Age-period-cohort analysis: an illustration of the problems in assessing interaction in one observation per cell data." *Communications in Statistics.Theory and Methods* 12 (23): 2779-2807.

Kupper, Lawrence L., Joseph Janis, Azza Karmous, and Bernard G. Greenberg. 1985. "Statistical Age-Period-Cohort Analysis: A Review and Critique." *Journal of Chronic Diseases* 38 (10):811-830.

Luo, Liying. 2013. Assessing Validity and Application Scope of the Intrinsic Estimator Approach to the Age-Period-Cohor Problem. Demography 50:

Mason, Karen Oppenheim, William M. Mason, H.H. Winsborough, W. Kenneth Poole. 1973. "Some Methodological Issues in Cohort Analysis of Archival Data." *American Sociological Review* 38 (2):242-258.

McCullough, P. and Nelder, J.A. 1983 *Generalized Linear Models*. New York: Chapman and Hall.

Raferty, Adrian E. 1995. Bayesian model selection in social research. *Sociological Methodology* 25:111–164.

Rodgers, Willard L. 1982a. "Estimable Functions of Age, Period, and Cohort Effects." *American Sociological Review* 47 (6):774-787.

Tu, Y-K, N. Kramer, and W-C Lee. 2012. Addressing the identification problem in age-period-cohort analysis: a tutorial on the use of partial least squares and principal components analysis. *Epidemiology* 23:583-593.

Yang, Yang, Wenjiang J. Fu, and Kenneth C. Land. 2004. "A Methodological Comparison of Age-Period-Cohort Models: The Intrinsic Estimator and Conventional Generalized Linear Models." *Sociological Methodology* 34 (1):75-110.

Yang, Yang. 2008. "Trends in U.S. Adult Chronic Disease Mortality, 1960-1999: Age, Period, and Cohort Variations." *Demography* 45 (2):387-416.

Yang, Yang, Sam Schulhofer-Wohl, Wenjiang J. Fu, Kenneth C. Land. 2008. "The Intrinsic Estimator for Age-Period-Cohort Analysis: What it is and how to use it." *American Journal of Sociology* 113 (6):1697-1736.